On the Inherent Segment Length in Music

Jensen, Karl Kristoffer

Published in:
Machine Audition

DOI (link to publication from Publisher):
10.4018/978-1-61520-919-4.ch013

Publication date:
2011

Document Version
Early version, also known as pre-print

Chapter 13
On the Inherent Segment Length in Music

Kristoffer Jensen
Aalborg University Esbjerg, Denmark

ABSTRACT

In this work, automatic segmentation is done using different original representations of music, corresponding to rhythm, chroma and timbre, and by calculating a shortest path through the selfsimilarity calculated from each time/feature representation. By varying the cost of inserting new segments, shorter segments, corresponding to grouping, or longer, corresponding to form, can be recognized. Each segmentation scale quality is analyzed through the use of the mean silhouette value. This permits automatic segmentation on different time scales and it gives indication on the inherent segment sizes in the music analyzed. Different methods are employed to verify the quality of the inherent segment sizes, by comparing them to the literature (grouping, chunks), by comparing them among themselves, and by measuring the strength of the inherent segment sizes.

INTRODUCTION

Music consists of sounds organized in time. These sounds can be understood from a rhythmic, timbral, or harmonic point of view, and they can be understood on different time scales, going from the very short (note onsets) to the medium (grouping), to the large scale with musical form. Note onsets, grouping and form are common musical terms, which can be compared to different aspects of audition, memory and grouping behavior. These terms can be compared to chunks, riffs, and other temporal segmentation terms currently used in music.

When identifying chunks, riffs, sections, forms, or other structural elements, do they really exist, or does the identification process create them? This work presents a method, based on automatic segmentation, that identifies the inherent structure sizes in music, i.e. gives indications as to what are the optimal segmentation sizes in the music. This work has implications for rhythmical and

DOI: 10.4018/978-1-61520-919-4.ch013

classical music understanding, and processing. Structure is a necessary dimension in most, if not all music, and if this structure should be made visible for any purpose, the methods presented here can help identifying the optimal structure. While this fundamental research gives a method for finding the optimal segment size in music, and results using this method, more work is needed in order to assess the inherent structure with certainty for all music. Until then, research and development of automatic segmentation of music should possibly ascertain the inherent structure in the music genres that is the aim of the work, prior to performing the segmentation.

Any feature, that can be calculated from the acoustics of the music, can be presented in a manner, for instance by taking the time-derivative, so as to give indication of the local changes in the music. Such an existence of a local change is not a guarantee of an inherent structure, however. In order to assess the quality of the segmentation, the relative distance (or any measure of similarity) within a segment should be compared to the distance to the other segments. If the segment is well grouped, and far, in some sense, to the other segments, then it is a good segmentation. A method for assessing the segmentation is the silhouette (Kaufman & Rousseeuw 1990). Given a segmentation, the mean of the silhouette value for all segments is a good measure of the quality of the segmentation. Therefore, if all possible segmentations are calculated, the associated mean silhouette values can be used to ascertain the best, i.e. the inherent structure sizes.

As to the question of which feature is used for temporal perception of music, Scheirer (1998) determined in several analysis by synthesis experiments that rhythm could not be perceived by amplitude alone, but needed some frequency dependent information, which he constructed using six band-pass filters. Several other studies have investigated the influence of timbre on structure. McAuley & Ayala (2002) found that timbre did not affect the recognition of familiar melodies, but that it had importance enough to hurt recognition on non-familiar melodies. McAdams (2002) studied contemporary and tonal music, and found that the orchestration affects the perceived similarity of musical segments strongly in some cases. He also found that musically trained listeners find structure through surface features (linked to the instrumentation) whereas untrained listeners focused on more abstract features (melodic contour, rhythm).

Deliège and Mélen (1997) postulates that music is segmented into sections of varying length using cue abstraction mechanism, and the principle of sameness and difference, and that the organization of the segmentation, reiterated at different hierarchical levels, permits the structure to be grasped. The cues (essentially motifs in classical music, and acoustic, instrumental, or temporal otherwise) act as reference points during long time spans. Deliège and Mélen furthermore illustrate this cue abstraction process through several experiments, finding, among other things, that musicians are more sensitive to structural functions, and that the structuring process is used for remembering, in particular, the first and last segment. In order to ensure that at least part of the full dimensionality of music is taken into account in the work presented here, three different features are used. One feature is believed to be related to tempo and rhythm, and it is called the rhythmogram. Another feature is considered related to the timbre perception, at least the time-varying perceptive spectrum, and it is called the timbregram. Finally, another feature is related to the note values in the music, and it is called chromagram. By using three features with distinctly different content, it is the aim to further assess the results on inherent and optimal segment size presented here.

Segmentation of music is often done for thumbnailing (music summary) purposes. This is supposedly a means for presenting music, prior to selling it, for instance in online stores. Other uses of segmentation are artistic, for instance for live mixing of music, for faster navigation,

where the knowledge of structural elements can be used for skipping similar elements, or related to music identification. Finally, segmentation automatic labeling of music can be beneficial for music analysis.

As for the methods for automatic segmentation, Foote (2000) introduced the use of selfsimilarity matrices, by convolving the selfsimilarity matrix with a checker kernel, thus calculating the novelty measure, which gives indications of the degree of novelty over time. Bartsch and Wakefield (2001) used the chroma representation for audio thumbnailing, by selecting the maximum of the time-lag matrix, which is the selfsimilarity matrix filtered along the diagonal in order to reveal similarities along extended regions of the songs. Goto (2003) and Chai & Vercoe (2003) also identify repeating segments using chroma representation, Goto (2003) using a similarity measure of the distance between vectors together with a method for integrating vectors into segments, and Chai & Vercoe (2003) by identifying local minima in the dynamic programming, which is an indicator of repetition of segments. Paulus and Klapuri (2008) in addition use Markov models to assign labels (Chorus/Verse, etc) to the segments.

In this work, focus will be on determining if there exists an inherent segment size in the music. Indeed, most segmentation methods are able to compute segmentation at different time scales, while the chosen segmentation size is left to the application development stage. Knowing the inherent time scale in music is done in the following manner. First, the feature estimation is presented, then the segmentation using dynamic programming is performed for all time scale, then a measure of the quality of the segmentation is calculated, and the peaks of this measure are identified and used as an indicator of the inherent segmentation size. Several methods for assessing the importance of the optimum segment sizes are employed and discussed in the conclusions.

FEATURE ESTIMATION

In order to perform a good segmentation of the songs, a robust feature is needed. Indeed, the feature used for segmentation can change the segmentation result significantly. Three different features are investigated here; the rhythmic feature (the rhythmogram, Jensen 2005) is based on the autocorrelation of the perceptual spectral flux (PSF, Jensen 2005). The PSF has high energy in the time position where perceptually important sound components, such as notes, have been introduced. The timbre feature (the timbregram) is based on the perceptual linear prediction (PLP), a speech front-end (Hermansky 1990), and the harmony feature (the chromagram) is based on the chroma (Bartsch & Wakefield 2001), calculated on the short-time Fourier transform (STFT). The Gaussian weighted spectrogram (GWS) is performed in order to improve resilience to noise and independence on block size for the timbregram and chromagram. A speech front-end, such as the PLP alters the STFT data by scaling the intensity and frequency so that it corresponds to the way the human auditory system perceives sounds. The chroma maps the energy of the FFT into twelve bands, corresponding to the twelve notes of one octave. By using the rhythmic, timbral, and harmonic features to identify the structure of the music, some of the different aspects of music perception are believed to be taken into account. More information of the feature estimation used here can be found in (Jensen 2007).

Rhythmogram

Any model of rhythm should have as basis some kind of feature that reacts to the note onsets. The note onsets mark the main characteristics of the rhythm. In a previous work (Jensen 2005), a large number of features were compared to an annotated database of twelve songs, and the perceptual spectral flux (PSF) was found to perform best. The PSF is calculated with a step size of 10 mil-

liseconds, and the block size of 46 milliseconds. As the spectral flux in the PSF is weighted so as to correspond roughly to the equal loudness contour, both low frequency sounds, such as bass drum, and high frequency sounds, such as hi-hat are equally well taken into account.

This frequency weighting is obtained in this work by a simple equal loudness contour model. The power function is introduced in order to simulate the intensity-loudness power law and reduce the random amplitude variations. These two steps are inspired from the PLP front-end (Hermansky 1990) used in speech recognition. The PSF was compared to other note onset detection features with good results on the percussive case in a recent study (Collins 2005). In order to obtain a more robust rhythm feature, the autocorrelation of the feature is now calculated on overlapping blocks of 8 seconds, with half a second step size (2 Hz feature sample rate). Only the information between zero and two seconds is retained. The autocorrelation is normalized so that the autocorrelation at zero lag equals one. If visualized with lag time on the y-axis, time position on the x-axis, and the autocorrelation values visualized as intensities, it gives a fast overview of the rhythmic evolution of a song. This representation, called rhythmogram (Jensen 2005), provides information about the rhythm and the evolution of the rhythm in time. The autocorrelation has been chosen instead of the fast Fourier transform FFT, for two reasons. First, it is believed to be more in accordance with the human perception of rhythm (Desain 1992), and second, it is believed to be more easily understood visually. The rhythmogram firstly gives information about the tempo of the song, along with the strength of the tempo, and secondly gives information about the time signature, although this information is not always clearly visible.

Timbregram

The timbre is understood here as the spectral estimate and done here using the perceptual linear prediction, PLP (Hermansky 1990). This involves using the bark (Sekey & Hanson 1984) scale, together with an amplitude scaling that gives an approximation of the human auditory system. The PLP is calculated with a block size of approximately 46 milliseconds and with a step size of 10 millisecond. The timbregram is a feature that is believed to capture orchestration of the music, mainly. In the timbregram, information about which instruments are participating in the music at the current time step is given, along with indications of what dynamic level the instruments are played. It represents the perceptual frequency axis in 25 steps. When an instrument is introduced in the music, it is often visible in the timbregram. It can also show the overall frequency content, i.e. older music lacks in bass and treble, pop music generally has energy on all frequencies, while some dance music (techno) only has energy in the treble and bass regions. The timbregram also reveals when sections are repeated, and in particular when sections are climaxed, with stronger instruments throughout. This is reflected with stronger values in the particular bark/time locations.

Chromagram

Note estimation is notoriously error-prone even if a lot of progress is done in the domain currently. There exists one estimate that is robust and related to the note values, the chroma, which is used here. In the chroma, only the relative content of energy in the twelve notes of the octave is found. No information of the octave of the notes is included. The chroma is calculated from the STFT, using a blocksize of 46 milliseconds and a stepsize of 10 milliseconds. The chroma is obtained by summing the energy of all peaks of 12 $\log_2$ of the frequencies having multiples of 12. The chromagram gives information about the note value, without information about the octave. This is a rather good measure of which chords are played, and also of the musical scale and tonality. If several notes are played for a moment, then this is clearly visible

in the chromagram. Also when a note is dropped, and another note is instead played more, this is also clearly reflected in the chromagram values.

Gaussian Windowed Spectrogram

If the raw features are used, it has been found that the detailed information sometimes overshadows the long-term changes in the music. If the features are calculated on short segments (10 to 50 milliseconds), they give detailed information in time, too varying to be used in the segmentation method used here. Instead, the features are calculated on a large segment, but localized in time by using the average of many STFT blocks multiplied with a Gaussian. This is called the Gaussian Weighted Spectragram, GWS. Using the GWS, all segments are used at all time steps, but the current block values are weighted higher than the more distant blocks. By averaging, using the Gaussian average, no specific time localization information is obtained of the individual notes or chords, but instead a general value of the time area is given. In this work, the averaging is done corresponding to a −3 dB window of approximately 1 second. After the GWS, the timbregram and chromagram has a stepsize of ½ second.

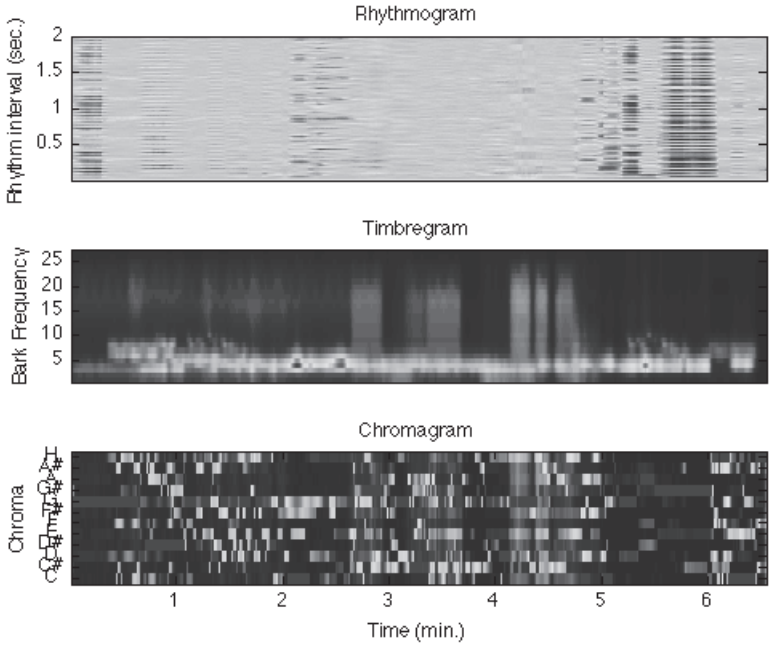
As an example of the features, the rhythmogram, timbregram and chromagram of August Engkilde – Beautiful Noise (Brumtone, 2008) is shown in Figure 1. All three features seem informative, although they do not give similar information. While the rhythm evolution is illustrated in the rhythmogram, it is the evolution of the timbre that is shown with the timbregram and the evolution of the note values that can be seen in the chromagram. Beautiful Noise is not a typical rhythmic piece of music, as can be seen from the lack of clear rhythm information in large part of the music. While the rhythmogram values are normalized, but instead the low time-lag values are set to zero. As these are also the autocorrelation value, which is by definition set to one, the relative strength of the higher time-lag correlations are reflected in the rhythmogram. In the case of Beautiful Noise, the very fast, almost vibrating rhythms have very similar repetitions, which are reflected as stronger values in the corresponding time segments in the rhythmogram. The timbregram reveals that this song has energy in two distinct frequency ranges, one low and one high, until almost three minutes. Then the high frequency component (a noise, windy sound) disappears. The timbregram is not normalized, so the crescendos are visible at a little after two minutes, at two and a half minute, and after five minutes. The chromagram is normalized, and reveals a single note played at the time, through this song. It changes from ‘G’ to ‘D#’, and back at around one minute, and to other note values elsewhere in the song. Both the rhythmogram, the timbregram, and the chromagram give pertinent information about the evolution in time of the music, and it seems judicious to investigate all three here.

SEGMENTATION

Automatic segmentation using dynamic programming has been proposed previously (Foote 2000, Bartsch & Wakefield 2001, Goto 2003, Chai 2003, Jensen 2005, Jensen et al 2005, Jensen 2007). In an automatic segmentation task, adjacent blocks are grouped together, forming segments. This can for instance correspond to the chorus/verse structure found in most rhythmic music, or to changes in the rhythmic pattern, in the orchestration or in the notes played.

The dynamic programming used here is based on the shortest-path algorithm (Cormen *et al* 2001) and done on self-similarity matrices, created from the original features (rhythm, chroma or timbre, Jensen 2007) by calculating the L2 norm of each time vector compared to all other time vectors, using a sequence of N vectors of each song that should be divided into a number of segments.

Figure 1. Rhythmogram (top), timbregram, and chromagram (bottom) of Beautiful Noise

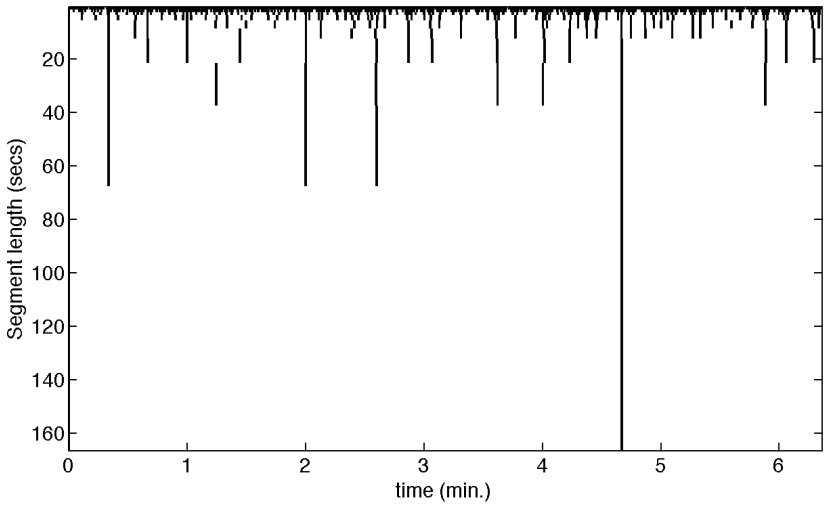


First, let the *cost c(i,j)* of a segment from block *i* to *j* be the weighted sum of the self-similarity and the cost of a new segment be a fixed cost α. Secondly, in order to compute a best possible segmentation, an edge-weighted directed graph *G* is constructed with the set of nodes being all the block of the song. For each possible segment an edge exists. The weight of the edge is $\alpha+c(i, j)$. A path in *G* from node *1* to node *N+1* corresponds to a complete segmentation, where each edge identifies the individual segment. The weight of the path is equal to the total cost of the corresponding segmentation. Therefore, a shortest path (or path with minimum total weight) from node *1* to node *N+1* gives a segmentation with minimum total cost. Such a shortest path can be computed in time $O(N^2)$.

The dynamic programming will cluster the time vectors into segments, as long as the vectors are similar. By varying the insertion cost α of new segments, segment boundaries can be found at different time scales. A low insertion cost will create boundaries corresponding to micro-level chunks, while a high insertion cost will only create few meso-level chunks. Thus, the same segmentation method can create segments of varying size, from short to long, from the grouping to the form of the music. An example of the segmentation is shown in Figure 2 for August Engkilde – Beautiful Noise (Brumtone 2008). As the segment cost (α) is increased, less and less segments are created, which in turn gives longer mean segment lengths.

The comparison of the segmentation done using the method presented here based on the three features rhythmogram, timbregram and chromagram reveals a F1 value of approximately 0.6 (Jensen 2007), corresponding to a matching recall and precision value of 50-70%. The comparison to manual segmentation gives F1 values slightly higher, at approximately 0.7 (Jensen 2007). This is an indication that the manual segmentation is done using different rhythmic, timbral and chroma cues, as the features are better matched to the manual segmentation than among themselves. Therefore,

Figure 2. Segmentation using the timbregram feature for all segment costs for August Engkilde – Beautiful Noise



it seems that all features should be employed in a segmentation task. As the final segmentation is not the target goal, this has not been deemed important here.

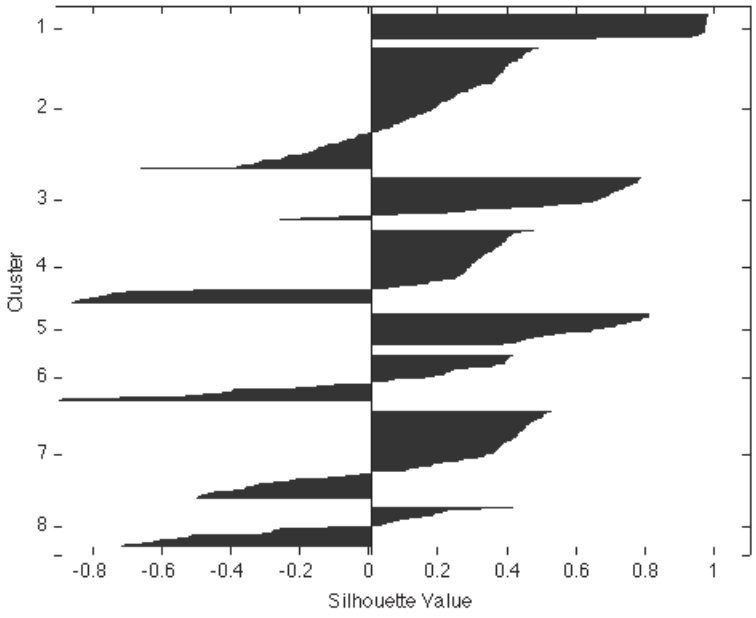
While many methods for segmentation of music exist, the problem of finding the inherent number of segments still exists. Kuhl related this to the notion of chunks. According to him, the chunk is an important element of music. A chunk is a short segment of a limited number of sound elements; a chunk consists of a beginning, a focal point (peak) and an ending. Kuhl (2007) extends the chunks to include microstructure (below 1/2 sec), mesostructure (chunks, the present, appr. 3-5 secs) and macrostructure (Superchunks, Kuhl and Jensen 2008) (at 30-40 secs).

BEST SEGMENT SIZE

The question investigated here is about whether there exist an inherent segment size in the music, and if so, if it is the same for different music, and if it is related to the chunk theory sizes. This question has been analyzed from different points-of-view in the literature. Huron (1996) investigated the melodic arch, and found a single arch up to 11 notes melodies, while melodies consisting of 12 or more notes present a double arch. This is, of course, related to the short-time memory theory of 7±2 (Miller 1956), but it does not give information about the time, only the number of notes.

In order to investigate this further, a database of varied music has been collected, and segmented using the shortest-path algorithm with the rhythm, timbre and chroma related parameters. The free variable, α, is varied in order to produce segment sizes between one block to the full song, i.e. all possible segment sizes. The classical way of investigating the clustering quality has to do with comparing the inter distance (the size of each cluster) to the extra distance (the distance between the clusters). Unfortunately, this cannot be computed for the one-cluster solution or the one cluster for each block solution, and it generally produces a 'U'-shape solution, with best values for small or large cluster sizes. A robust estimate of the cluster quality is the silhouette (Kaufman & Rousseeuw 1990), calculated for each observation i as

Figure 3. Silhouette plot for Beautiful Noise. A high positive values indicate that the observation is well clustered, while a large negative values indicate a bad clustering for the observation.



$$s = \frac{b_i - a_i}{\max(a_i, b_i)}, \qquad (1)$$

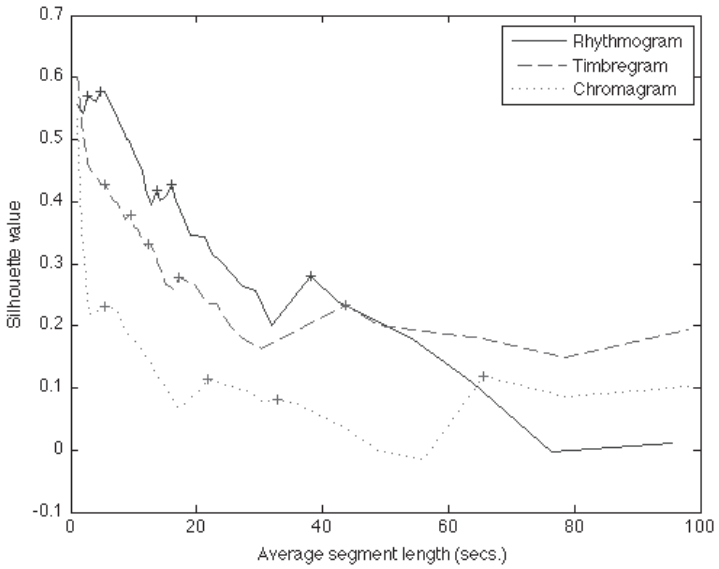
where *a* is the average dissimilarity to all other points in its own cluster and *b* is the minimum of the average dissimilarities of *i* to all objects in another cluster. The silhouette value for each observation is always comprised between -1 and 1. If the silhouette value is large, i.e. close to 1, the observation is centered in the cluster, while if the value is low, the observation is also close to other clusters. The silhouette for a clustering solution can be calculated as the average of each observations silhouette value. An example of the silhouette for the segmentation using the timbregram feature of Beautiful Noise is shown in Figure 3. There are eight clusters with an average length of 49 seconds. The average silhouette value is 0.2. Some of the clusters, the first and fifth in particular, have high silhouette values throughout, while some of the others have negative silhouette values for some of the observations of the cluster.

Segmentation Analysis

The mean silhouette value is now calculated for each new segmentation cost in order to analyze the quality of the different segmentations. One silhouette value is retained for each average cluster length, calculated as the total length divided by the number of clusters. An example of the mean silhouette value as a function of average cluster size for Beautiful Noise is shown in Figure 4. The average silhouette is plotted for the rhythmogram, timbregram and chromagram, along with indications of the peaks in each silhouette plot using plus '+' signs. This song has silhouette peaks for different segment lengths, including for the chunk size at approximately 5 seconds, and the superchunk size at approximately 40 seconds. Other optimum chunk sizes are also visible.

Nine songs of classical, pop/rap and jazz genres have been segmented using the rhythmogram, timbregram and chromagram features for varying new segment cost. Each song has a number of silhouette

Figure 4. Mean silhouette value as a function of average segment length for Beautiful Noise. Peaks are indicated with plus '+' signs.

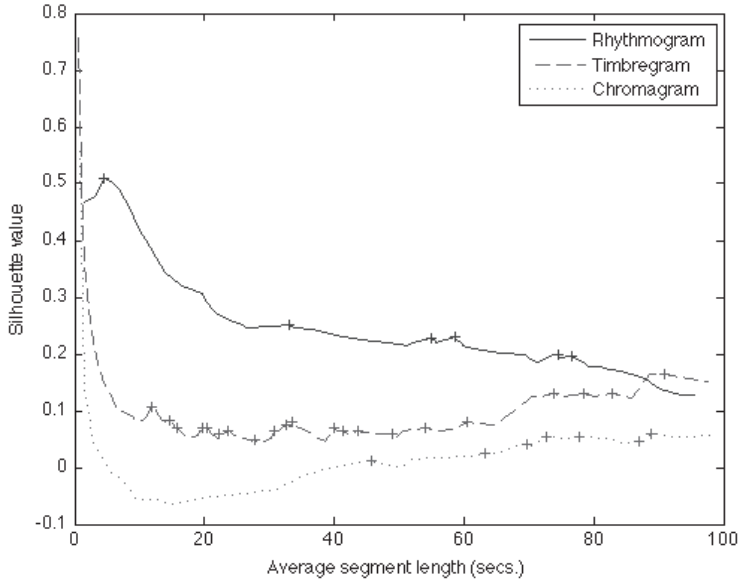


peaks for each feature, giving an indication of the inherent segment size as a function of the average segment length. Visually, the rhythmogram seems to give a better result. To compare to the chunk theory of Kühl (2007), the peaks of the average silhouette values can be identifies as belonging to the range 3-5 seconds, corresponding to the chunk size, and between 30-40 seconds, corresponding to the superchunk size. For the nine songs there has been found 5 (19.23%) chunk matches for rhythmogram, 8 (30.77%) superchunk matches, 4 (8.7%) chunk matches for timbregram, 5 (10.87%) superchunk matches and 2 (9.52%) chunk matches for chromagram, 4 (19.05%) superchunk matches. The rhythmogram performs significantly better than the other features for this particular task. This is also visible, if the mean of the nine songs is calculated and plotted (Figure 5). Indeed, the rhythmogram silhouette plot presents a prominent peak at the chunk level, around 5 second average segment length, and also one peak at approximately 30 seconds average segment length. The timbregram and chromagram has a more 'U' shaped silhouette value, effectively preventing any silhouette peak at the chunk level. Several other possible peak positions also exist, for instance around 60 seconds, and around 80 seconds.

Another distinction that can be made is between the short-term memory and the long-term memory. Snyder (2000) relates the short-term memory to melodic and rhythmic grouping and situates it between 1/16 second to 8 second, and the long-term memory to musical form, and situates this above 8 seconds. If the question is; what is more prominent, grouping or form, then the study performed here gives indications that form is most prominent, as there is 10 (38.46%), 10 (21.74%) and 4 (19.05%) peaks below 8 seconds (corresponding to grouping) for rhythm, timbre and chroma, respectively. Thus grouping is seemingly more related to rhythm, and less to timbre and chroma.

As to the question of the similarity of the peak position of the silhouette as a function of average segment length, the normalized histogram of the segmentation peaks are calculated, along with a measure of the peakedness of each peak. The normalized histogram gives values of the

Figure 5. The mean silhouette values for nine songs and rhythmogram, timbregram and chromagram



relative occurrences of different optimal segment lengths. The peakedness is calculated as the relative strength of the peak, divided by the width of the two surrounding samples

$$p = \frac{s_i}{4 \cdot (s_{i-1} + s_{i+1}) \cdot (l_{i+1} - l_{i-1})}, \qquad (2)$$

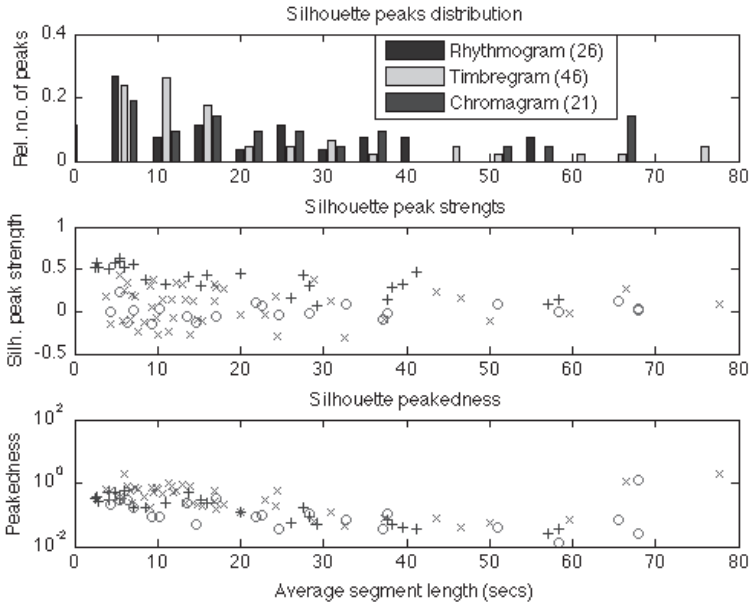
where s_i is the silhouette value at index i, and l_i is the average length at index i. This normalization by the width is necessary, as the silhouette values are not uniformly sampled along the average segment length axis. In addition, the peak silhouette value is also retained. The silhouette peak distrubution, values and peakedness for nine songs are shown in Figure 6.

There are 26 rhythmogram silhouette peaks, 46 timbregram, and 21 chromagram peaks in all in the nine songs. In the histogram (Figure 6, top), the rhythmogram have an apparent peak at approximately 5, 15, 25, 37, and 55 seconds average segment length, the timbregram has peaks at 10, 30, 45, and 75 seconds, and the chromagram has peaks at 5, 15, 35 and 65 seconds average segment length. In the peak value subplot (middle), the rhythmogram silhouette values seem higher than the other values. The mean silhouette peak values are 0.45, 0.15, and -0.03, showing a significantly better value for the rhythmogram, and a rather unusable value for the chromagram. As for the peakedness (Figure 6, bottom), the peakedness values are decreasing with the average length of the segments, except for a few peaks with high peakedness values at very high segment lengths. The timbregram has apparently the highest peakedness values, and the chromagram the lowest. The mean of the peakedness is 0.23, 0.64, and 0.17 for rhythmogram, timbregram, and chromagram, respectively. However, the actual peak values are deemed more important, and they are showing the rhythmogram to be the best feature for segmentation.

ACTUAL INHERENT SEGMENT BOUNDARIES

Given the method presented here, it is now possible to identify the optimal segmentation sizes.

Figure 6. Histogram of silhouette peak position (top), the peak silhouette values (middle), and the silhouette peakedness (bottom) as function for average segment length for nine songs. The rhythmogram values are depicted with a '+', the timbregram a 'x', and the chromagram a 'o' in the middle and lower subplot.

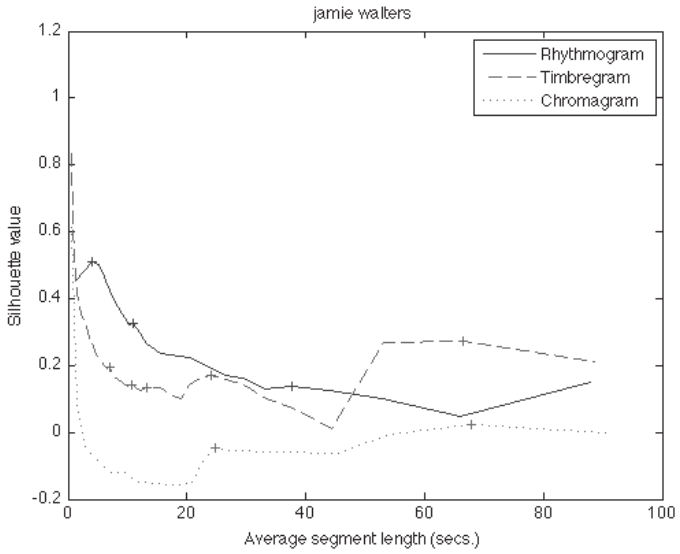


The question is now, what do these sizes represent in the music. As an example, the song Hold On by Jamie Walters (Atlantic 1994) is further investigated. The mean silhouette values as function of average segmentation length for this song is shown in Figure 7. As for the other songs, the rhythmogram mean silhouette values have a rising peak for short average segment lengths, and then decreasing, while both the timbregram and the chromagram-based silhouette values have a 'U'-shape, i.e. the silhouette values decrease to a minimum, and then rise again, with only local maxima. The maxima of silhouette for rhythmogram are found at approximately 4, 11 and 38 seconds, while the timbregram silhouette maxima are found at (0.5), 7, 11, 13, 24 and 66, while for the chromagram, the maxima of the silhouette are found at (0.5), 25 and 68. The (0.5) seconds are the peaks at the cases where all observations have a separate segment (all observations a grouped into individual segments), and thus the average segment length is equal to the sampling rate of the features.

The musigram plots along with the automatic segmentation boundaries for the same song are found in Figure 8. The rhythmogram is shown in the upper subplot, the timbregram in the middle subplot and the chromagram in the lower subplot. The ensemble is called musigram (Kuhl & Jensen 2008). The rhythmogram reveals alternating sections with more or less strong pulse. The timbregram reveals a weak intro and first sections, a stronger section, which is repeated (1min30, and 2min30), and possibly repeated in a crescendo at 3min10. Similar observations can be made in the chromagram. However, the segmentations found using the automatic segmentation do not necessarily find the same segments, which is possibly impeding on the quality of these experiments. However, it is not believed to be very influential in the results of the experiments.

The average segment lengths for the automatic segmentation boundaries in Figure 8 are 11, 13

Figure 7. Mean silhouette value as a function of average segments length for Hold On – Jamie Walter (1994)



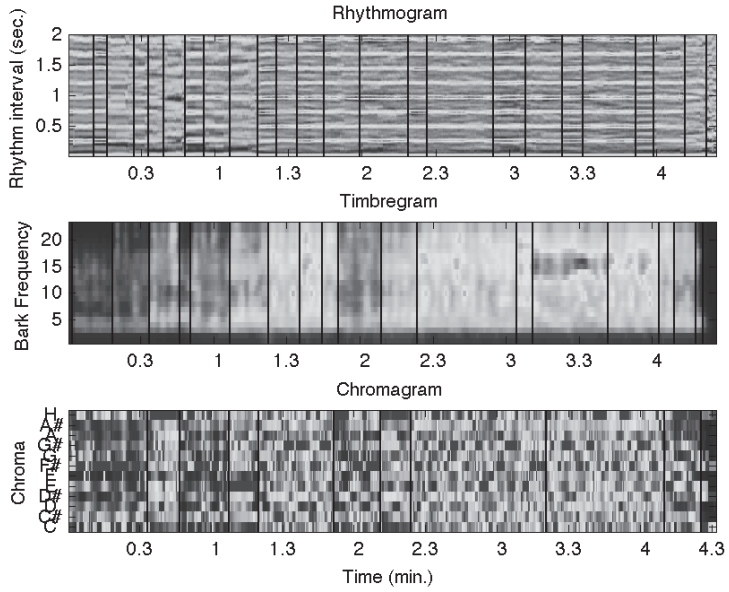
and 25 seconds. It is clear, from the analysis of more songs, that the automatic segmentation gives at the same time shorter and longer segments. For these average segmentation lengths, the rhythmogram gives segment lengths between 5.5 and 27 seconds (standard deviation is 5.35), the timbregram renders segment lengths between 1.5 and 41 (std=9.7 seconds), and the chromagram segment lengths between 12.5 and 56.5 seconds (std=15.81). It is therefore difficult to say whether the optimum segment lengths correspond or not to other theories, such as the chunk theory of Kuhl (2007). However, indications towards such a correspondence is nonetheless observed. First, there is often peaks in the silhouette values for different segment lengths, which corresponds somewhat to the micro, meso and super chunks of Kuhl, which has sizes at 0.5, 3-5 and 30-40 seconds. However, often, the segmentation based on the different features renders different optimal segmentation lengths using the silhouette method.

Album Study

On the question of how reproducible the results of the study of inherent segment sizes in music, a second experiment has been performed. Are the inherent segment sizes similar across the music genres, or within a music style, or are the inherent segment sizes changing to a degree if there are no systematic values to be found? In order to investigate this further, a full album, The Beatles – Sgt Peppers Lonely Hearts Club Band (Parlophone/Capitol, 1967) has been analyzed in the same manner as above. The three features, rhythmogram, timbregram and chromagram have been calculated from the acoustics of each song of the album, then the automatic segmentation has been done for all possible segment sizes. Finally, the silhouette values have been calculated as a measure of the quality of each segment size. If these silhouette values have peaks on similar segment sizes for the different songs, this is an indication that the inherent segment sizes are similar across this particular album, which can be seen as a sample of a genre. If the silhouette peaks are scattered around, the inherent segment sizes are individual for each song.

In order to investigate this, the relative number of silhouette peaks for different segment sizes has been calculated for the Beatles album, along with

Figure 8. Musigram (rhythmogram (top), timbregram, and chromagram (bottom), along with automatic segmentation boundaries (illustrated with vertical lines) obtained using each feature for Hold On - Jamie Walter (1994)

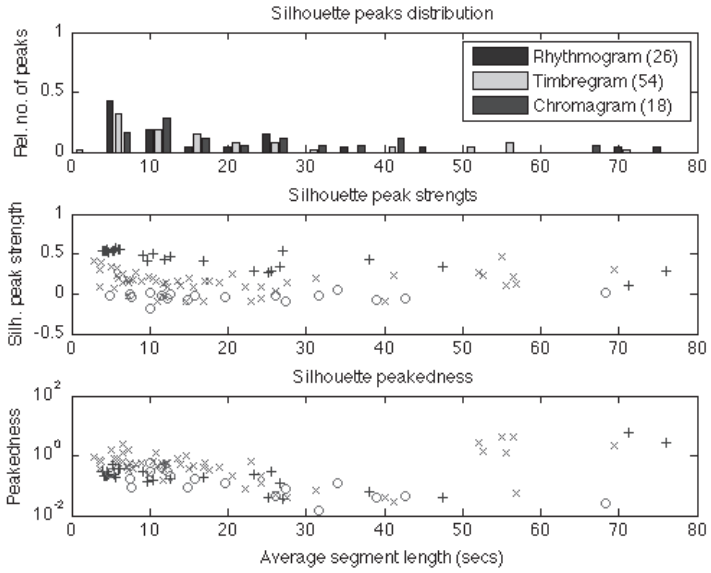


the peak values, and the peakedness values, and are shown in Figure 9.

The timbregram and chromagram always renders a peak at the shortest possible segment size (0.5 second). This has not been taken into account here. When compared with the similar data for the nine songs of varied genres (Figure 6), the differences are seemingly minor. The rhythmogram renders peaks at 5, 25 and 72 seconds, but not at 15 and 37 seconds. The timbregram renders peaks at 5, 25 and 55 seconds, and the chromagram peaks at 10, 25, 400 and 65 seconds. The rhythmogram silhouette peak values are significantly higher than the timbregram values, which are significantly higher than the chromgram values, with average silhouette peak strength of 0.45, 0.15, and -0.03 for rhythmgoram, timbregram and chromagram. As for the nine songs of varying genres, only the rhythmogram has acceptable silhouette values. When analyzing the silhouette peakedness, the same decreasing with average segment length peakedness is observed for the Beatles album as for the nine songs, along with some high peakedness values for the vey high segment lengths. The mean peakedness values are 0.53, 0.79, and 0.17 for rhythmogram, timbregram, and chromagram, respectively. The timbregram again has the best peakedness, and the chromagram the worst. All in all, however, the rhythmogram performs best, with better silhouette values, and also there is a distinctive grouping of the rhythmogram peaks in Figure 9. This is more visible in the peak values and peakedness subplots, than in the histogram. There is one group consisting of 11 observations between four and six seconds, and one group with five observations between 11 and 13 seconds, and another group with five observations between 23 and 27 seconds. All in all, these three groups account for 21 of the 26 peaks for the rhythmogram.

As for the chunk and superchunk identification, there has been found 10 (38.46%) chunks and 5 (19.23%) superchunk for the rhythmogram, 8 (14.81%) chunk and 5 (9.26%) superchunk for

Figure 9. Histogram of silhouette peaks as a function of average segment lengths for The Beatles - Sgt Peppers Lonely Hearts Club Band (top), and silhouette peakedness values for same (bottom). The rhythmogram values are depicted with a '+', the timbregram a 'x', and the chromagram a 'o' in the middle and lower subplot.



timbregram, and 1 (5.56%) chunk and 6 (33.33%) superchunk for the chromagram.

There has been found 11 (42.31%), 17 (31.48%) and 3 (16.67%) below 8 seconds (grouping) for rhythm, timbre and chroma, respectively. These numbers confirm that the rhythmogram corresponds better to the chunk and grouping theory for the Beatles songs, as it did for the nine songs of varied genres above.

CONCLUSION

Researchers and developers have found an increasing number of reasons for segmenting music into smaller segments, be it for thumbnailing, re-mixing (artistic), identification, playlist generation or other music information retrieval purposes, copyright issues, music analysis purposes, or yet other issues. However, while segmentation can easily be done, a grounded theory of the obtained, or wished for segmentation sizes should be available. This can be found in the music theory, for instance by the grouping theory of Lerdahl & Jackendorff (1973). However, their subdivision into groups of notes (often corresponding to measures) are based on rules, of which one states that groups at the same level have the same duration, which is not found in this work. While they state the rules as *preferences* that can be based on melodic or rhythmic proximity and continuity, it does not seem consistent with the results obtained here. Results from memory research (Snyder 2000) can also be used as the ground reference. Snyder refers to echoic memory (early processes) for event fusion, where fundamental units are formed by comparison with 0.25 seconds, the short-term memory for melodic and rhythmic grouping (by comparison up to 8 seconds), and long-term memory for formal sectioning by comparison up to one hour. Snyder (2000) relates this to the Gestalt theory grouping mechanisms of proxim-

ity (events close in time or pitch will be grouped together. Proximity is the primary grouping force at the melodic and rhythmic level (Snyder 2000, p 40). The second factor in grouping is similarity (events judged as similar, mainly with respect to timbre, will be grouped together). A third factor is continuity (events change in the same direction, for instance pitch). These grouping mechanisms give rise to closure, that can operate at the grouping level, or the phrase level, which is the largest group the short-term memory can handle. When several grouping mechanisms occur at the same time, intensification occurs, which gives rise to higher-level grouping. Other higher-level grouping mechanisms are parallelism (repeated smaller groups), or recurrence of pitch. The higher-level grouping demands long-term memory and they operate at a higher level in the brain, as compared to the smaller time-scale grouping. The higher-level grouping is learned while the shorter grouping is not. Snyder (2000) further divides the higher level grouping into the objective set, which is related to a particular music, and the subjective set, which is related to a style of music. Both sets are learned by listening to the music repeatedly. Snyder (2000) also related the shorter grouping to the 7±2 theory (Miller 1956), that states that the short-term memory can remember between five to nine elements.

While the music theory and memory-based research can give grounded results for the segmentation tasks, they are seemingly not giving the full truth. The music theory operated with a constant size of segments, which is not what is observed by automatic segmentation. Obviously, both the music theory and the memory-based grouping are related in many senses, which Snyder (2000) also points out. These works find some of its basis in traditional music theory and solfège. The main problem with these theories is the seemingly lack of emphasis on the actual sound, the timbre, and the performance with respect to timing and dynamics, in particular.

In contrast to these theories, the work presented here only takes into account the acoustics of the music. There is no high-level music theory, and no prior understanding based on the mechanisms of the brain.

Automatic segmentation is performed here by calculating the shortest-path through the selfsimilarity of different audio features, which can be related to rhythm, timbre and chroma. By varying the cost of inserting a new segment, different time scales are created, going from the short (seconds) to the long (up to 100 seconds). The question that is investigated here is whether music has an inherent segment length. Indeed, both music theory and brain research have theories about different time scale, which is visible in the music scores, and in different psycho-physical experiment regarding memory. In order to investigate the inherent time scale in the music, the silhouette values are calculated for all observations (blocks) for each segmentation scale. The mean of the silhouette values is a good measure of the quality of the segmentation. By matching the peaks of the silhouette values to the chunk theory of 3-5 seconds and the superchunk theory of 30-40 seconds, a measure of the inherent segmentation size has been found, together with indications of which feature that permits a better analysis of this. The rhythmogram has the best match for the chunk (19%) and superchunk (31%) levels, respectively for nine songs of varying genres, while a Beatles album gives 38%, and 19% for rhythmogram. In this case, the chromagram has better superchunk result (33%). Visually inspecting the mean silhouette values of nine songs of varying genres further reveals that only the rhythmogram has a peak at the chunk level, while all three features has peaks at the vicinity of the superchunk size of 40 seconds. Most peaks are situated above the average length of eight seconds, which gives indication that form is more prominent than grouping in the songs investigated here. Indications that form is more prominent than grouping (Snyder 2000) is given, along with indications that grouping is

more prominent in the rhythmic features than in the timbral or chroma features. Further analysis of the silhouette peaks reveals the systematic occurrences of peaks in several average segment length positions, including short segments around 5 seconds, medium length segments at around 20 seconds, and longer segments. These findings are similar for a collection of nine songs of varying genres, and the Sgt Peppers album of the Beatles. The rhythmogram has systematically larger silhouette values at the peaks, with an average of approximately 0.4, while the timbregram and chromagram have mean silhouette peak values of 0.1 and 0, which is an indication that all observations could just as well belong to another segment than the one they belong to. An analysis of the peakedness of the silhouette peaks reveals that the timbregram produces stronger peaks, and chromagram the weaker peaks. Seemingly, the low length peaks have higher peakedness than the peaks found for longer segments.

While more work is necessary in order to confirm the findings here, several indicative conclusions can nonetheless be drawn; 1) The music investigated here, which is of varying genres, has inherent segment sizes of different length. These inherent segment lengths are found for all songs with relative small variations. 2) The rhythmogram performs best when the found inherent segment lengths are compared to theory, and it is also the only feature that has an acceptable average silhouette peak value.

REFERENCES

Bartsch, M. A. & Wakefield, G.H. (2001). To Catch a Chorus: Using Chroma-Based Representations For Audio Thumbnailing. *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics* (CD). New York: IEEE.

Chai, W., & Vercoe, B. (2003). Music thumbnailing via structural analysis. *Proceedings of ACM Multimedia Conference*. November.

Collins, N. (2005). A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. *Proceedings of AES 118th Convention*, Barcelona, Spain, May.

Cormen, T. H., Stein, C., Rivest, R. L., & Leiserson, C. E. (2001). *Introduction to Algorithms*,(2nd Ed.). Boston: MIT Press. and New York: McGraw-Hill.

Deliege, I., & Melen, P. (1997). Cue abstraction in the representation of musical form. In Deliège, J. Sloboda (Eds). *Perception and cognition of music(*387-412).East Sussex, England: Psychology Press.

Desain, P. (1992). A (de)composable theory of rhythm. *Music Perception*, *9*(4), 439–454.

Foote, J. (2000). Automatic Audio Segmentation using a Measure of Audio Novelty. [July.]. *Proceedings of IEEE International Conference on Multimedia and Expo*, *1*, 452–455.

Goto, M. (2003). A chorus-section detecting method for musical audio signals. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, *437-440*(April).

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*(4), 1738–1752. doi:10.1121/1.399423

Huron, D. (1996). The Melodic Arch in Western Folk songs. *Computing in Musicology*, *10*, 323.

Jensen, K. (2005). A causal rhythm grouping. Proceedings of 2nd International Symposium on Computer Music Modeling and Retrieval (CMMR '04),Denmark (LNCS, vol. 3310, pp. 83-95)

Jensen, K. (2007). Multiple scale music segmentation using rhythm, timbre and harmony. *EURASIP Journal on Applied Signal Processing, Special issue on Music Information Retrieval Based on Signal Processing*.

Jensen, K., Xu, J., & Zachariasen, M. (2005). Rhythm-based segmentation of Popular Chinese Music. *Proceeding of the ISMIR*. London, UK, (pp.374-380).

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kühl, O. (2007). *Musical Semantics*. Bern: Peter Lang.

Kuhl, O., & Jensen, K. (2008). *Retrieving and recreating Musical Form. Lectures Notes in Computer Science*. New York: Springer-Verlag.

Lerdahl, F., & Jackendoff, J. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

McAdams, S. (2002). Musical similarity and dynamic processing in musical context. *Proceedings of the ISMA* (CD). Mexico City, Mexico.

McAuley, J. D., & Ayala, C. (2002). The effect of timbre on melody recognition by familiarity. *Meeting of the A.S.A.*, Cancun, Mexico (abstract).

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two. *Psychological Review*, *63*, 81–97. doi:10.1037/h0043158

Paulus, J., & Klapuri, A. (2008). Labelling the Structural Parts of a Music Piece with Markov Models. *Proceedings of the 2008 Computers in Music Modeling and Retrieval*, (pp.137-147), Copenhagen, Denmark.

Scheirer, E. (1998). Tempo and Beat Analysis of Acoustic Musical Signals. *The Journal of the Acoustical Society of America*, *103*(1), 588–601. doi:10.1121/1.421129

Sekey, A., & Hanson, B. A. (1984). Improved 1-bark bandwidth auditory filter. *The Journal of the Acoustical Society of America*, *75*(6), 1902–1904. doi:10.1121/1.390954

Snyder, B. (2000). *Music and Memory. An Introduction*. Cambridge, Mass.: The MIT Press.